DESIGN AND DEVELOPMENT OF A MACHINE LEARNING FRAMEWORK FOR AUTOMATIC QUESTION GENERATION

Synopsis

Submitted in the partial fulfilment for the requirement of

Doctor of Philosophy

In

Physics and Computer Science

Submitted by

Monika Singh

Under the supervision of

Prof. C. Vasantha Lakshmi
(Co-Supervisor)
Department of Physics and Computer Science
Faculty of Science

Prof. C. Patvardhan
(Supervisor)
Department of Electrical Engineering
Faculty of Engineering

DEPARTMENT OF PHYSICS AND COMPUTER SCIENCE, FACULTY OF SCIENCE

DAYALBAGH EDUCATIONAL INSTITUTE (DEEMED TO BE UNIVERSITY)

DAYALBAGH, AGRA-282005

2022

1. Introduction

Curiosity is the nature of humans. They interrogate and try to be more knowledgeable by interacting with their environment. Usually, no conversation is complete without question answering. In classrooms, students learn more by asking several questions from their teachers. Thus, asking questions plays an important role in their learning process. Similarly, teachers can evaluate students' performance by asking them a variety of questions. Automatic Question-generation helps learners in identifying their knowledge gap. Tutors face problems in producing good hints to motivate students [1] [2] [3].They need to ask relevant questions to assess the students' knowledge and address their knowledge deficiency. The users who access different search engines like Google, Yahoo, Yandex, etc. tend to input a few words instead of a full sentence or in the form of a query [4] [5]. Thus, to retrieve answers quickly, reformulation of guided questions is needed. IBM Watson, a question-answering computer system performs self-training with the help of an automatic of the question answering system [6]. This system also automates the answering process so that human experts are not needed to define the ground truth answers manually. The question generation process helps the question-answering model to decide whether the question is answerable or not. Rather than giving wrong answer, this process abstains from answering. Further, automation of question generation helps in data augmentation to refine the question-answering model by automatically generating unanswerable questions[7]. Thus, from regular conversations to verbal tutoring and evaluation, questions are a primary source of learning.

Many automated question generation models make use of standard datasets for training. These datasets are traditionally collected from crowdsourcing by generating question-answer pairs manually. The question-answering system can generate a large volume of Question-answer pairs from raw and unstructured documents, which helps in generating refined, large, and strong question answering datasets automatically [8]. SQuAD (Stanford Question Answering Dataset) is a reading comprehension dataset in which questions and answers are manually generated by crowdworkers on a set of Wikipedia articles. It contains 100,000+ question-answer pairs on 500+ articles.However, manually curating a large dataset for a new language rather than English of comparable size like the

SQuAD dataset is tedious and expensive.Existing datasets can be leveraged with a Question-generation model to generate a dataset for a new language [9].

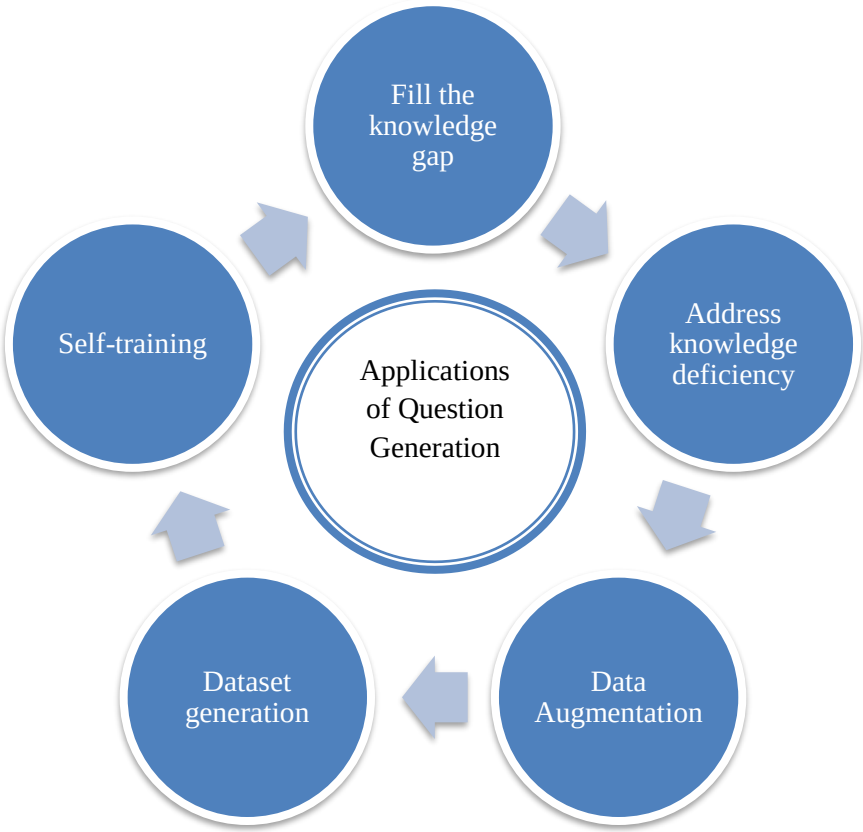

Figure 1: Applications of Question Generation.

2. Classification of questions

A variety of questions are used in various applications as shown in Figure 1. For a general question-answer system, the questions are classified in many different ways as follows.

2.1. Function-based classification of questions

Questions can be classified according to Function as [10].

Table 1: Function-based question classification

S.No.	Question Type	Question Description	Answer Description	Example
1.	Fact-based Questions	Questions based on general facts	Short phrase answers	Who is the Prime Minister of India?
2.	List-based Questions	Questions that are asked for the list within answers	A single phrase or phrase having proper explanations	List five species of birds found in India?

3.	Reason-based Questions	Questions that are asked for explanations and opinions	Summery having enough comprehensive explanations and a variety of opinions	What are your views on rising food insecurity?
4.	Solution-based Questions	Questions that are asked for problem shooting	The answer should follow the logical order to give a proper explanation	How is a patient given treatment by a doctor?
5.	Definition-based Questions	Questions that are asked for the illustration of concepts	The answer should be a summary	What does the bar graph talk about?
6.	Navigation-based Questions	Questions are asked for finding resources or websites	Answers are in the form of a URL	Where can I download the SQuAD dataset?

2.2. Reasoning-based classification of questions

Based on the reasoning, the questions are classified as follows.

2.2.1. Single-hop reasoning questions

When reasoning is performed within a single document or paragraph it is termed as single-hop reasoning. Questions generated on these single documents or paragraphs are referred to as single-hop questions. For example, the SQuAD dataset is a single hop reasoning question-answering dataset.

2.2.2. Multi-hop reasoning questions

Reasoning over multiple documents is multi-hop reasoning. Questions generated on these multiple paragraphs or documents with multiple supporting facts for question-generation are referred to as a multi-hop questions. For example, the HotPotQA dataset is a multi-hop reasoning question-answering dataset.

2.3. Questions classification based on Bloom's Taxonomy

Bloom's Taxonomy was generated by Benjamin Bloom in 1956, where educational learning objectives are classified into levels of complexity and specificity [11].

Table 2: Bloom's Taxonomy based question classification

TAXONOMY	EXPLANATION	SAMPLE QUESTION
Create	Production of original or new work. (assemble, conjecture, formulate, investigate, design, construct, develop, author)	Write the summary of chapter one of a novel in your own words
Evaluate	Justify a decision (appraise, defend, select, value, weigh,	Outline Mohan's plan to join the community. Then, evaluate the

	argue, judge, support, critique)	quality of the scheme with its strengths and weaknesses
Analyze	Make connections between ideas. (organize, compare, distinguish, experiment, test, differentiate, relate, contrast, examine, question)	What motivates Mohan to join the community? Use textual evidence to support your answer
Apply	Use the information in a new situation. (Implement, use, interpret, schedule, execute, solve, demonstrate, operate, sketch)	Using Mohan's knowledge about others, make a prediction that what happens when Mohan goes to the restaurant
Understand	Describe concepts or ideas. (Describe, explain, locate, report, translate, classify, discuss, identify, recognize, select)	Why is the achievement of Rashmi is more prestigious than others?
Remember	Memorize facts and basic concepts. (Duplicate, memorize, state, define, list, repeat)	At what age were people got married?

According to Bloom's Taxonomy, questions are of two types, higher-order reading questions and lower-order reading questions.

2.3.1. Higher-order reading questions

Questions that require implicit knowledge, longer than few phrases, cannot acquire answers directly from the document, may have more than one right answer, and require multiple ideas to analyze and evaluate the document or create something new are higher-order reading questions. The questions that fall under the top three levels of Bloom’s Taxonomy are classified as higher-order reading questions.

2.3.2. Lower order reading questions

Questions that are directly drawn from the document and have a quicker response as compared to higher-order reading questions are lower-order reading questions. Questions that do not fall under the top three levels of this Taxonomy are classified as lower-order reading questions.

2.4. Classification of Questions based on usage

Table 3: Usage based question classification

S. No.	Types	Explanation
1.	The Dichotomous Questions	These are normally Yes or No closed-ended questions for basic verification.
2.	Multiple Choice Questions	In multiple-choice questions, questions are provided with a list of options and respondents must choose one or more responses from those options. The options contain the right answers or answers, close alternatives, incorrect answers, distractors. Here single select questions are represented by the radio

		button and multi-choice questions are represented by checkboxes.
3.	Rank Order Scaling Questions	Rank order allows giving rank preference to the respondent in a question in order of question type choice. These types of questions help in understanding the weight of each option given by the respondents. Drag and drop question type is one of the categories of rank order scaling question where rearrangement of options can be done by respondents based on their importance.
4.	Text Slider Questions	These are rating scale questions in which an interactive slider is used in answer to choose the most suitable option. The scale of the option is well defined and on the same constancy. It is used to measure the intensity and direction of an approach.
5.	Likert Scale Questions	The Likert scale is the most used tool by market researchers to evaluate the viewpoint and attitudes of their target audiences towards a given subject. A five, seven, or nine-point agreement scale is used to check the respondent's agreement with numerous statements. This scale can be unipolar or bipolar.
6.	Semantic Differential Scale	This type of question scale is used to rate company, entity, brand, and products within a multi-point rating option.
7.	Stapel Scale Questions	Stapel Scale Questionsare used to collect the respondent's preferences about a particular subject. Questions contain even response options.
8.	Constant Sum Questions	These are alsorank order type questions where options are in the form of numeric.
9.	Comment Box Open-Ended Questions	For collecting any feedback or any suggestions that can be exceptionally long as well as short, the comment box open-ended questionsare used. In this the responder can answer depending on knowledge, feeling, and knowledge, so they are open text formatted.
10.	Text Questions	A text question and a comment box are remarkably similar, but in the Text Question, the entered data are generally validated and controlled. The Text Questionsare generally divided into three sub-types: Single row text, Numeric Textbox, Email address.
11.	Contact Information Questions	Contact Information Questionsareopen-ended questions. These have multiple rows of text which are indicated with a title. In this, the entered text is regulated.These kinds of questions gather information or detail of the responder such as name, email address, contact number, address, etc.
12.	Demographic Questions	The question which is used to record the demographic data extracted from the set of the population is known as Demographic Questions. The gender, age, number of children, income, race, etc. are identified by Demographic Questions. To give more correct and accurate information about the group of people, Demographic Questionsare used.

13.	Matrix Table Questions	These types of questionsare ordered in a format of a table in which the questions are listed on the left side while the options of the answers are on the top of the table. In the multipoint matrix table, the radio buttons are given to choose the answer for multiple aspects of any question. In the multi-select matrix table, there is a checkbox to choose or select the answer in question.
14.	Side-by-Side Matrix Questions	Side-by-Side Matrix Questionsare used for organizing surveys so that the importance and the level of satisfaction with the services which are offered to the user can be recognized. It provides multiple options for rating simultaneously. Our strengths and areas in which the improvementsare needed are easily identified by Side-by-Side Matrix Questions.
15.	Star Rating Questions	For displaying feelings and emotions or to rank attributes, the odd number of stars are used. These types of rating questionsare known as the Star Rating Questions. Theyprovide a neutral point or satisfaction in a rating survey. If the number of stars is higher, then the agreement toward the statement is higher.
16.	Max Diff Questions	The Max Diff Questionsare used for indicating the worst and best attributes from the set of attributes. They can have only one of the options as a final response.
17.	Push to Social Questions	These types of questionsare used for allowing us to share our reviews or feedback on social networking sites such as Facebook, Google+, Twitter, etc. These are used to make the popularity of your brand as well as to address customer's issues on social media.
18.	Visual Analog Questions	These are close-ended questions andcan be managed in two types- An odd point smiley question type- In this, there is a Five-point rating scale in which a visual representative nature is given to rate. Thumbs Up/Thumbs Down question type- It is a two-point question in which the respondent gives the feedback either a positive or a negative response.
19.	Image Questions	Image Questionsare the better technique to increase the rate of response. In this, the options are given as the images
20.	Net Promoter Score (NPS) Questions	Net Promoter Score (NPS) Questionsarequestion types to get the compact feedback related to the satisfaction level of the customer about our organization or brand. In this, the response to the feedback can be measured on a scale of 1-10.
21.	Van Westendorp-Price Sensitivity Questions	These are used by market researchers to scale the consumer's thought of the value of services and products. They help in the scope of adjusting the

		price and offerings.
22.	Date/Time Questions	The Date/Time questionsare used to gather the information related to the Date and Time filled by the respondent.
23.	CAPTCHA Questions	The CAPTCHA Questions limit the fake and false responses during the surveys or data collection by using automated computer programs.
24.	Calendar Questions	The Calendar Questions provide the feature of taking Date and time in the format of the calendar as an input to the respondent.
25.	Interactive Maps Questions	They permit the respondent or responder to click on the location on a map by hovering over the locations on the map or by clicking on it. When the respondent scrolls a map then the selection of the locations can be made by getting a pop-up by hovering on it.
26.	Lookup Table Questions	These Lookup tables complete the information automatically in a systematic way. These are the mixture of a single-line editable textbox and a drop-down list.
27.	Store Locator Questions	These types of questions help in finding or locating the store for purpose of a business near a given area or address.
28.	TubePulse Questions	These types of questions permit the responder to give feedback about a particular topic in a discussion. In this, the provocative point can be represented as a video.
29.	Reference Data Questions	These Reference Data Questionsareused to validate the zip codes for a given country against those that are stored in the database.

The questions starting with 'when, why, who, what, where are the Wh-questions. These are the most prominent and important question types for acquiring knowledge and gathering information in a particular domain. Some question-answering system developed wh-questions for reading comprehension [12]-[13]. Similarly, Varga, generates ‘how many’ type questions in addition to prior wh-questions [14].

All the research work developed so far mostly generates factoid questions. Many question-generation (QG) systems use different kinds of key-phrase extraction techniques to generate factoid questions [15]-[16]-[17]. Some QG system uses multiple supporting facts for generating high-quality multi-hop factoid questions by applying Multi-Hop Question Generation Model [18]. Some system uses an answer span annotation as a fact to generate high-quality questions with the help of GPT-2 and the pre-trained BERT model [20]. Wasim, Mahmood, Asim, and Ghani, improve the factoid and list type of questions in the Biomedical question answering [21]. A tool used for the pedagogical system

should aid the students in such a way that they become capable enough to apply their learning in a new situation and this can be achieved by analyzing deep and shallow questions. There are many QG systems that generate deep and shallow questions [22]-[23]-[24].

Gap-filling questions and multiple-choice questions are widely used for taking the assessments in various fields as they are easy to conduct and easy to evaluate, There exist some QG systems that generate such questions by applying various pattern matching approaches [25]-[26].

3. Question Generation Approaches

The question generation approaches are classified into two classes: rule-based question generation and neural-based question generation. Rule-based question generation approaches include syntax-based question generation, template-based question generation, and semantics-based question generation, while neural-based question generation approaches use a sequence-to-sequence question generation algorithm. Rule-based techniques are summarized in Table 4 and Neural-based techniques are summarized in Table 5.

3.1. Template-based Question Generation

Preferable templates for question-generation are created by humans and for each answer phrase, pertinent templates are used. This approach is advised for specific motive applications. Question templates like when-would-Y-happen, why-Y, what-would-happen-if, and what-would-happen-when are used in question-generation for text having conditional context. In the above question templates, Y is the placeholder mapped to various semantic roles labeled by a semantic role labeler. Here the templates which involve 'why' are used for generating linguistic modality questions and those which involve 'when' are used for generating temporal questions. After classification of an individual's citation into classes like the result, aim, application, opinion, system, and method, a suitable question-template is used by an AQG (Automatic Question Generation) system [27][28].

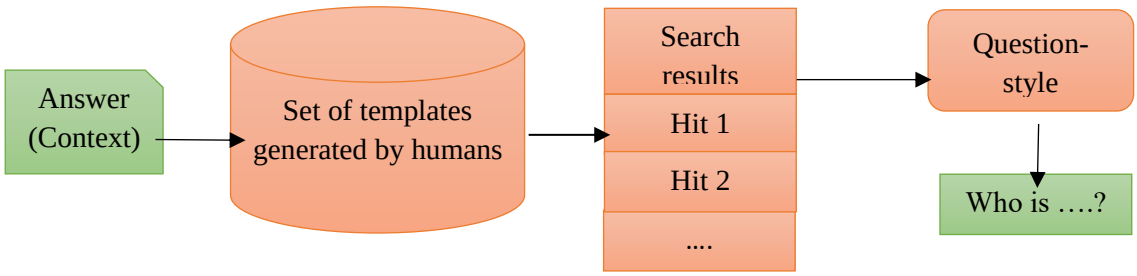

Figure 2: Template-based Question Generation pipeline

In automatic question generation, each QG approach has its own advantages for generating different types of questions. E. Riloff and M. Thelen presented a template-based approach that used multiple views of text: lexical and semantic clues to select the answer worthy text in the passage [12]. N.-T. Le, T. Kojiri and N. Pinkwart used transformation templates to generate semantic questions [29]. M. Liu, R. A. Calvo and V. Rus focused on automatic question generation by applying different preprocessing stages like Tregex expression rules and citation classification to extract meaningful sentences from the text, and then used predefined templates for question generation [16]. K. Mazidi and R. D. Nielsen proposed a lexical-semantic tree, and templates were manually created according to this QG tree [30].

3.2. Syntax-based question generation

Syntax-based question generation approaches are suitable for short sentences where explicit factual information of a sentence is used to create questions. This approach follows three steps to generate questions: (a) deletion of the target concept, (b) Placement of a suitable question keyword at the initial position of the question, and (c) conversion of the verb into correct grammatical form. For example, the sentence Dr. A. P. J. Abdul Kalam served as the 11^{th} President of India from 2002 to 2007 can be transformed into: who served as the 11^{th} President of India from 2002 to 2007? [28][27]

The syntax-based approaches provide better coverage of the text for question generation. P. Mannem, R. Prasad and A. Joshi presented parser for content selection, then selected target content (for question generation) by using mandatory arguments, optional arguments, and copular verbs with dependency parse, and then used appropriate rules for question generation [13]. S. Pal, T. Mondal, P. Pakray, D. Das and S. Bandyopadhyay determined the cue phrase in the clause. This cue phrase was replaced by a specified question word for question generation. In these clauses, reordering the chunks generated more possible questions based on rule-based grammatical knowledge [31].

3.3. Semantics-based question generation

Semantics-based question generation involves background knowledge and semantic information of a statement / topic to generate questions. The semantic information could be retrieved from diverse sources, such as ARK (Archival Resource Key) and WordNet (a syntax-based tool that is used to generate questions from English sentences). This approach typically involves three steps: (a) Question templates generation without WordNet, (b) question construction using retrieved hyponyms and

previously generated templates, and (c) Utilize example sentences from WordNet and ARK to generate final version of questions [28][27].

Background resources are used by a semantic-based approach to generate more interesting questions. K. Mazidi generated conceptual and descriptive questions by applying Natural Language Understanding (NLU) techniques like Topic modeling from the R topic modeling package, Noun phrase extraction from Python package TextBlob, Terminology extraction from WordNet, and selecting section topic as a noun phrase [32]. O. Keklik, T. Tuglular and S. Tekir generated deep and shallow questions by using the number of parsers for POS tagging (SENNA's POS tagging algorithm), Chunking (SENNA's chunking algorithm), Dependency parsing (SpaCy's algorithm), Named entity recognition (SpaCy's algorithm) [24].

3.4. Neural-based question generation

Rule-based approach does not generate high-level questions and needs human attention to generate manual rules [27]. Thus, to overcome all the infirmity of the rule-based approaches, neural-based approaches were presented.

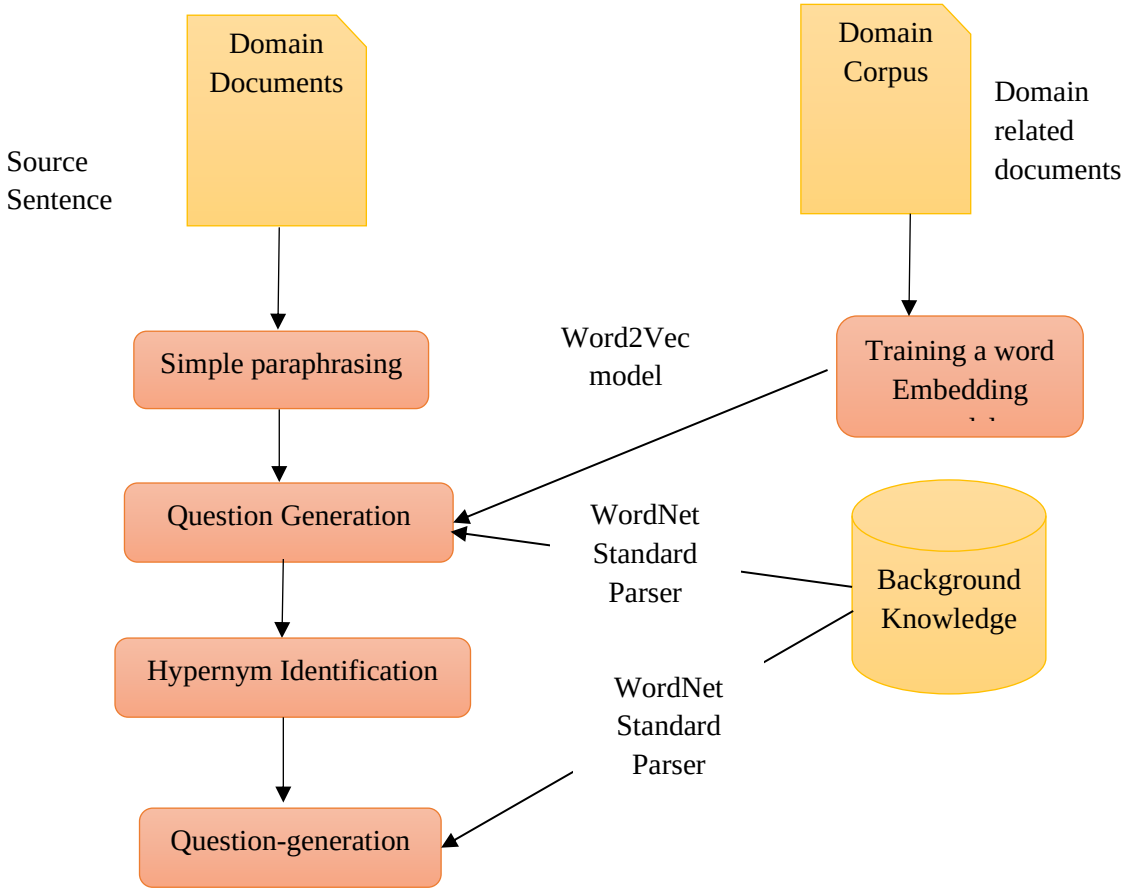

Fig. 2: Template-based Question Generation pipeline.

Neural-based question generation is purely a data-driven sequence to a sequence learning algorithm that directly converts selected text passages into well-formed questions in which both the input text and output questions are considered as two distinct sequences. This approach does not rely on just memorizing the learned material and focuses on understanding. The sequence-to-sequence learning algorithm takes the previously selected tokens and an input text and generates the next token by maximizing the conditional log-likelihood of the predicted question sequence. Long short-term memory encoder-decoder architecture is used for training the corpus to calculate the conditional probability of each token.

More recently, a transformer-based model has been proposed to perform QG. This model provides the power of Attention to refer specific points within the text paragraph. Additionally, decoders have extra encoder-decoder attention layer that are stacked on top of each other to focus on the important part of the input sequence. Sequence-to-sequence based models are complex, requires extra features and additional mechanism. However, to overcome this infirmities, transformer-based models are used [63] [64].

4. Question Evaluation Techniques

There are some famous manual and automatic evaluation models to evaluate the quality of the generated questions [33].

4.1. Manual Evaluation

In Manual Evaluation, human evaluators are employed for evaluating system performance. This is a simple and easily available technique. Slow speed and expensive human resources are some limitations of this method [28].

i. Extrinsic Evaluation

Extrinsic evaluation focuses on the question's contribution to the performance of other external systems. It measures the learning gain and efficiency of users in terms of time and their effectiveness. It supports human participation. For example, this evaluation measures how well generated questions help the users to improve their search efficiency by comparing them with other query suggestion services (Yahoo or google assistant) [4].

ii. Summative Evaluation

Summative evaluation is applied after completion of an AQG. It informs user about the defined goals that have been achieved by the final version of the system. It involves participation of human evaluators [34].

Unfortunately, manual evaluation is not good and has some limitations: it provides inconsistent results, and is expensive and slow. Even after adopting standard best practices in study, human beings are inconsistent to make decision about what is good. In addition, human evaluators need to be recruited and trained. All these factors indicate that manual evaluations entail large investment of resources and time consuming. Thus, to get consistent results within less time, automatic evaluation methods are favored by most of the researchers [34].

4.2. Automatic Evaluation

Automatic evaluation involves the development of algorithms to mimic the behavior of the human evaluators and compare the generated output with the gold standard [34].

i. Intrinsic Evaluation

Intrinsic Evaluation evaluates the functionality of AQG against benchmark results. It involves Automatic Question Generation (AQG) evaluation based on: Precision and Recall score [15][35] and conciseness, grammaticality or other quality metrics [28][4].

Precision

Precision is the percentage of relevant questions among the top N questions. It is calculated as:

Qaqg: Questions generated by AQG,

Qmanually: Questions generated manually.

$$\text{Precision} = (Q_{aqg} \cap Q_{manually})/Q_{aqg}$$

Recall

Recall is the percentage of the relevant questions that were retrieved. It is calculated as:

$$\text{Recall} = (Q_{aqg} \cap Q_{manually})/Q_{manually}$$

ii. Formative Evaluation

Formative Evaluation is applied during the development of an AQG. It informs the designer about progress of the intended goals [34].

iii. BLEU

BLEU metric (BiLingualEvaluation Understudy) is a precision-based metric used to evaluate machine translation [36]. Quality is checked by matching the similarity between the human-generated questions and the machine's output. If machine translation is closer to human translation, generated questions have better quality. BLEU is an incredibly famous and inexpensive metric and claims that it has a high correlation with human opinions. BLEU's score lies between 0 and 1. It is calculated for a single translated sentence by comparing it with a reference translation of good quality and averaged on the whole corpus to get the overall quality. If the score is 1 that means the translated text is more similar to reference translation. Grammatical correctness cannot be checked by the BLEU metric [37][38][39][18][7].

iv. METEOR

METEOR metric (Metric for evaluation of translation with explicit ordering) is an F-measure metric used to evaluate machine translation and is based on unigrams [40]. Additional features of this metric are synonymy and stemming matching. It creates alignment by mapping the candidate's token with the token in the reference. If both tokens are synonyms or the same of their stem match, then they are aligned with each other. This alignment is calculated in the form of precision and recall values, and both are combined into the F-measure score in which recall has higher weightage [37][38][39][18][7].

v. ROUGE

ROUGE metric (Recall-Oriented Understudy for Gisting Evaluation) is an F-measure metric used to evaluate the automatic summarization and machine translation [41]. The automatically produced system's translation and summary are compared with the reference set [37][38][39][18][7].

Types of ROUGE evaluation metrics:

- ROUGE-N: ROUGE-1 for unigram and ROUGE-2 for bigram.
- ROUGE-L: Based on the Longest Common Subsequence (LCS).
- ROUGE-W: Weighted LCS based.
- ROUGE-S: Skip bigram-based.
- ROUGE-SU: Skip bigram and unigram-based.

vi. Sentence and Word Mover's Similarity

Sentence and Word Mover's Similarity is the earth mover's distance-based similarity that is used to evaluate system-generated summaries [42]. Here word embedding is weighted based on its frequency and sentence embedding is weighted based on its length. After this, a linear program is used to calculate the cumulative distance of the moving document's words and sentences with the reference documents.

vii. BERTScore

In the BERTScore evaluation metric, a similarity score is calculated for each token in the candidate sentences with those in the reference sentence by feeding them through BERT model independently [43]. Based on pairwise cosine similarity, an alignment between candidate and reference words is calculated. Use this alignment to aggregate the precision and recall values before aggregating with the F1 score. This metric performs better than human evaluation.

viii. Conditional BERTScore

Conditional BERTScore is an extension of BERTScore that incorporates the concatenation of context, question and answer described by BERT separator tokens as the input to BERT. Then, the BERT representations of the answer words are extracted and compute BERTScore. Thus, the representation of the answer words are conditioned with the context and question [43].

Table 4: A brief review of rule-based approach

Topic	Types of Questions	Question-generation	Question Evaluation	Reference
1. A rule-based question answering system for reading comprehension tests. (2000)	Wh-questions start with when, why, who, what, where.	Text preprocessing includes morphological analysis, semantic class tagging, part-of-speech tagging, and entity recognition. For question -generation Quarc uses lexical and semantic clues and creates a separate set of rules for each question type by rewarding four-point values: clue (+3), good_clue (+4), confident (+6), and slam_dunk (+20).	HumSent answers are answers judged by humans to be the best answers. Quarc achieves overall 40% accuracy on the HumSet answer key.	[12]
Automatic Generation of Prediction	Visualization questions in multiple-choice	It uses an MCode interpreter and interaction package for question	Human evaluation is performed. And it provides	[44]

Questions during Program Visualization. (2007)	form.	generation in Jeliot-3, a program visualization tool.	automation of question generation in the Jeliot-3 tool.	
3. Automation of Question Generation from Sentences. (2010)	Factoid questions.	Preprocessing step includes oak system, Part of speech tagger, Named entity tagger, and Charniak Parser. For question-generation 90 predefined interaction rules are used.	It uses recall and Precision scores for evaluation. Where the overall recall score is 0.315 and the overall precision score is 0.485 out of 1.	[15]
4. Question generation form paragraphs at UPenn: QGSTEC system description. (2010)	General questions, medium questions, specific questions.	Text preprocessing includes SRL parses and dependency parses. Identify the verb complex using dependency parses and apply syntactic transformation rules for question construction. Some other used tools are BART sentence splitter, MorphAdornerlemmatizer, Bidirectional POS tagger, Illinois named entity tagger, Bidirectional LTAG dependency parser, and ASSERT semantic role labeler.	Question ranking is performed by the depth of the predicate in the dependency parse and by the number of pronouns present in the question.	[13]
5. QGSTEC system description- JUQGG: a rule-based approach. (2010)	Questions start with what, which, whose, who, whom, when, where, and why.	Pre-processing of the corpus includes parsing through Semantic role labeler and Stanford dependency parser, named entity recognition, chunking through Conditional random field chunker, and casual relators. The question recognizer module is used to determine the cue phrase in the clause. This cue phrase is replaced by a specified question word for question generation. In these clauses, reordering the chunks generates more possible questions based on rule-based grammatical knowledge.	Human evaluators are hired to check the quality of the questions.	[31]
6. WLV: A question generation system for the QGSTEC 2010	Questions start with what, which, whose, who, whom, when, where,	It uses shallow parsing, sentence transformation, key phrase identification (with Charniak parser), named entity recognition,	Questions are rated by two humans by considering relevancy, syntactic correctness, and	[14]

Task B. (2010)	why, and how many.	and using language resources like WordNet, corpora, Wikipedia at preprocessing step. And generates different syntactic rules for different question types on subject-verb-object sentences with at least one key phrase.	fluency of the question. The system generates 115 questions out of 180 target questions and does not score very well in terms of relevancy and syntactic correctness measures.	
7. Question generation with minimal recursion semantics. (2010)	Deep questions (why, what-if, and how), shallow questions (what, where, how many, who, when, which, and yes/no).	Preprocessing includes Stanford named entity recognizer, an Ontology NE tagger, regular expression NE tagger, feature structure chart, and Minimal recursion semantic decomposition is done to convert the complex sentence into a shorter semantic form.Minimal recursion semantic transformation is performed to replace the named entity's EPS with (wh) question word EPS to generate suitable questions.	Precision grammar (ERG) is used to guaranteed the grammaticality of the outputs.	[22]
8. Automatic Question Generation for Educational Application- The State of the Art. (2014)	Semantic-based questions, Yes or No questions, Questions start with who, what, whom, where, which, and how.	Automatic question generation in the educational application includes two approaches: the transformation-based approach with some set of transformation rules and the template-based approach with some predefined questions templates.	The results are evaluated in the form of precision and recall. And template-based systems achieved great results as compared to transformation-based systems.	[29]
Automatic Generation and Ranking of Questions for Critical Review. (2014)	Factoid questions.	The preprocessing stage includes Tregex expression rules, citation classification consists of Naïve Bayes classifier. Questions are generated by some predefined rules. The ranking of the questions is performed by two approaches: pointwise approach logistic regression and pairwise approach RankSVM.	Human evaluators and Fisher's least significant difference (LSD) tests are employed for evaluation.	[16]
Linguistic Considerations in Automatic	Factoid questions.	SENNA software is used to divide the source into target sentences at preprocessing	Human evaluation is performed by calculating the	[17]

Question Generation. (2014)		step. Question generation follows a predefined generation pattern.	average score of the predicate position and comparing it with other baselines. The proposed method achieves a 44% reduction in error rate relative to the prior system and a 61% reduction in grammatical judgment.	
11. Leveraging multiple views of text for automatic question generation. (2015)	Composition questions, semantically-oriented questions, factoid questions, list questions, describe questions and the questions start with what, how, when, why, where, when.	Preprocessing includes Stanford dependency parser, predicate-argument parser from SENNA. Templates are manually created according to the QG tree. The subject-verb-object template creates what and who questions, templates based on discourse cues with ArgM modifiers form when, why, how, describe, and consequence questions. Templates that support multiple features from the QG tree form more semantically-oriented questions.	Human evaluation can be done by calculating: average scores on three scales (not acceptable, borderline acceptable, and acceptable) and achieves a 44% reduction in error rate as compared to state-of-the-art QG systems at that time.	[30]
12. Automatic Question Generation Based on Analysis of Sentence Structure. (2016)	Factual questions.	The pre-processing step includes splitting text into sentences, token annotation (to recognize POS tags and semantic meaning), and related token terms are merged into one. The template-based method is applied for question construction by taking structures of the sentence as patterns with information like POS tag, named entity tag, super sense tags obtained from WordNet dictionary or linked data categories, and using multiple patterns, in which searching is based on traversing through the hierarchy.	Evaluation is performed by experimenting: comparing the proposed approach with an existing basic approach by investigating the number of patterns on a piece of text. And the results show that they need only 27 patterns to search instead of 72 (from the basic approach).	[45]
13. Automatic Question	Deep and Shallow	Analyze the central semantic content for each	This study is evaluated by using	[23]

Generation from NLU to NLG. (2016)	questions.	independent clause for each sentence by gathering data from the SRL parse and dependency parse.50 question templates are matched with sentence patterns to generate higher quality questions.	Amazon's Mechanical Turk Serviceand achieves a 55% acceptability rate.	
14. Automatic Question Generation approaches and Evaluation techniques. (2017)	Gap-fill questions, Multiple choice questions, Factoid-based questions, Deep learning questions.	It includes Sentence simplification by using independent clause, appositive phrase, leading prepositional phrase, relative clause, discourse marker, noun participle, and answer phrase selection by identifying PP and noun as an answer phrase or by employing semantic role information and named entity recognition for answer phrase. Question generation includes a template-based approach, syntax-based approach, and semantic approach.	AQG evaluation techniques include intrinsic evaluation, extrinsic evaluation, black-box evaluation, glass-box evaluation, automatic evaluation, manual evaluation, formative evaluation, summative evaluation, and the bystander Turing test.	[28]
15. Automatic question generation from the passage. (2017)	Conceptual questions, descriptive questions, factual comprehension questions.	Applying NLU techniques like Topic modeling from the R topic modeling package, Noun phrase extraction from python package TextBlob, Terminology extraction from WordNet, and selecting section topic as a noun phrase. Questions are generated through appropriate methods by using the preprocessing stage.	Evaluation is performed in three ways: question breadth and depth, question linguistic quality, and Question importance.The linguistic quality of the proposed approach on paragraph-level is 16% better than on sentence-level and 29% over the Heilman & Smith system. It generates 8% more important questions as compared to sentence-level questions.	[32]
16. Good-quality question generation for academic support. (2018)	Fact-based questions, Gap Fill questions.	Any rule-based automatic question generation.An Automated Question Quality Enhancement (AQQE) module transforms the Ill-formed questions (IFQ) into Well-formed	174 questions are classified as WFQ or IFQ questions by human annotators and AQQE module (by multiple linear regression model)	[35]

		questions (WFQ) by predicting the scale of a question using a multiple linear regression model to classify whether the question is acceptable or not, if not then uses Stanford parser to parse, POS tagging and Tregex to identify and retrieve the nodes and finally use Tsurgeon tool to modify the irrelevant nodes.	having 87% precision score and 99% recall scores.	
17. Rule-based automatic question generation using semantic role labeling. (2019)	Deep and shallow questions.	Expand all the contractions and idiomatic language is eliminated. After that, the number of parsers for POS tagging (SENNA's POS tagging algorithm), Chunking (SENNA's chunking algorithm), Dependency parsing (SpaCy's algorithm), Named entity recognition (SpaCy's algorithm), and Semantic role labeling (AllenNLP's algorithm) are applied.Question construction uses some predefined rules or templates like dependency-based templates, NER-based templates, and SRL-based templates.	Evaluation metrics like BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, and ROUGE-L are used in which proposed rules outperform BLEU-2, METEOR, and ROUGE-L scores. In human evaluation, proposed rules produce the closest scores to the human-generated questions.	[24]
18.Automatic Question Generation. (2020)	Fill-in-the-blank questions. Factoid questions start with What, When, Where, and Who. High-level comprehension questions. Six types of questions for vocabulary assessment include synonym, hypernym, definition, antonym, and cloze questions.	Text preprocessing techniques includes Sentence simplification, Parsing, Named Entity Recognition (NER), Semantic Role Labeling (SRL), and anaphora resolution in which the state-of-the-art NER taggers are, Stanford NER and Illinois NER and SRL parsers are, ASSERT and SENNA. Existing AQG is classified into three categories: syntax-based, template-based, and semantics-based question generation approaches. Linguistic properties of the input sentences are analyzed by employing	AQG systems are evaluated manually by two or more domain experts along with several aspects of quality. Some automatic evaluation metrics are BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, and ROUGE-L.	[27]

		some core NLP tools such as Stanford Phrase Structure Parser, Tregex Tree Searching Language and Tool, Supersense Tagger, and ARKref Noun Phrase Coreference Tool. AQG approach is termed as a Neural-based approach that uses a sequence-to-sequence QG algorithm. Question-ranking can be done by the question-ranking module by combining the syntactic similarity scores and the topic relevance scores. Another used model is the statistical model, based on the least-squares linear regression function. One is using a machine learning approach for question ranking. Question ranking can also be done by the depth of the predicate in the dependency parse tree.		

Table 5: A brief review of the Neural-based approach

Title	Encoder and Decoder	Dataset	Evaluation	Reference
1. Learning to Ask: Neural Question Generation for Reading Comprehension. (2017)	Encoder: Attention-based Sentence and paragraph encoder with bidirectional LSTM network. Decoder: Decoder with LSTM network and a global attention mechanism.	The SQuAD dataset	BLEU, METEOR, and ROUGE-L metrics are used and Investigates that the sentence-level information generates more natural and higher-order reading questions as compared to paragraph-level information.	[46]
2.Identifying Where to Focus on Reading Comprehension for Neural Question Generation. (2017)	Encoder: First, important sentences are selected by applying to encode using convolution + maximum pooling operation. Then bidirectional LSTM network is used. Decoder: Global attention-based LSTM decoder.	The SQuAD dataset.	The evaluation system includes BLEU and METEOR metrics.The question generation system with an important sentence selection method outperforms the existing state-of-the-art approaches and generates factual questions.	[47]
3.Harvesting Paragraph-	Encoder: Four types of embedding include	The SQuAD	BLEU and METEOR metrics, and Binary	[8]

Level Question-Answer Pairs from Wikipedia. (2018)	Coreference Position Feature Embedding, Refined Coreference Position Feature Embedding, Answer Feature Embedding, and Word Embedding, andan Encoder with a bidirectional LSTM network. Decoder: Question Decoder with LSTM network and Attention and Copy based Mechanism.	dataset	Overlap, and Proportional Overlap metrics in addition to exact match are used for evaluation. And generates fact-based and higher-order answer-specific questions and produces a corpus of 1.26 million question-answer pairs.	
4.Paragraph-level Neural Question Generation with Maxout Pointer and Gated Self-attention Networks. (2018)	Encoder: RNN based Passage and answer encoder with answer tagging and gated self-attention network. Decoder: RNN based Decoder with attention and max out pointer mechanism.	The SQuAD dataset and the MS MARCO dataset.	BLEU, METEOR, and ROUGE-L evaluation metrics are used. The question generation system with an important sentence selection method outperforms the existing state-of-the-art approaches and generates high-level textual questions.	[48]
5.Putting the Horse Before the Cart: A Generator-Evaluator Framework for Question Generation from Text. (2019)	Encoder: Answer encoder with attention and boundary pointer network, and Sentence encoder with two-layer bidirectional LSTM network. Decoder: Question decoder with single-layer LSTM network, and copy and coverage mechanism.	The SQuAD Dataset.	BLEU, GLEU, and ROUGE-L and two new reward functions for the generated question and answer against the ground truth are used for evaluation. And generates textual fact-based, reason-based, higher-order, and syntactic and semantic questions.	[19]
6.Let's Ask Again: Refine Network for Automatic Question Generation. (2019)	Encoder: Passage and answer encoder include Embedding Layer with Glove model, Contextual Layer with a bidirectional-LSTM network, and Passage-Answer Fusion Layer. Decoder: Preliminary decoder with LSTM Network, and Refinement decoder with Dual Attention Network.	The SQuAD, HOTPOT , and DROP datasets	BLEU, ROUGE-L, METEOR, and QBLEU4 are used and defeats the existing state-of-the-art model by improving and refining the first stage of generated questions and generates lower-order, grammatically correct, and fluent questions.	[49]
7.Reinforced	Encoder: Rationale	The	BLEU and ROUGE-L	[50]

Dynamic Reasoning for Conversational Question Generation. (2019)	Selection, Encoding, and Reasoning with bidirectional LSTM network. Decoder: Dynamic Reasoning and Decoding with LSTM network.	CoQA dataset.	metrics are used. It generates high-quality, fact-based, and lower-order reading questions by focusing on the most relevant text span for generating the next question.	
8.Learning to Answer by Learning to Ask: Getting the Best of GPT-2 and BERT Worlds. (2019)	Encoder: Transformer Encoder BERT. Decoder: Transformer Decoder GPT-2 model.	The SQuAD dataset	BLEU metric, ROUGE-L metric, and the BERT QA model as a surrogate measure for assessing the quality of the question generation are used and generates the diverse, fact-based, and semantically correct questions with relevant answers that are present in the given passage.	[20]
9.Learning to Generate Questions with Adaptive Copying Neural Networks. (2019)	Encoder: Attention-based encoder with bi-directional LSTM network. Decoder: Copying mechanism-based decoder with attention-based LSTM network.	The SQuAD dataset.	Evaluation measures are BLEU-n and ROUGE-L metrics, and some baseline algorithms like Seq2Seq, Du-sent, Du-para. It overcomes the flaws of question generation at sentence and paragraph level by generating deep, shallow, and proper natural questions and perform state-of-the-art literature in terms of used evaluation metrics.	[51]
10.Learning to Ask Unanswerable Questions for Machine Reading Comprehension. (2019)	Encoder: It contains three types of embedding: the word embeddings, the character embeddings, and the token type embedding. And the Paragraph encoder with single-layer bi-directional recurrent neural networks and LSTM units. Decoder: Decoder with single-layer uni-directional LSTM network for generating factual and higher-order unanswerable questions.	The SQuAD dataset.	GLEU, BLEU, and ROUGE-L evaluation metrics are used for evaluation. It improves multiple reading comprehension and constructs training data by introducing a syntactic data augmentation technique and pair-to-sequence model and generates an unanswerable question. It achieves state-of-the-art literature in terms of used evaluation metrics.	[7]
11.Weak Supervision Enhanced Generative	Encoder: Relation Guider consists of word embedding, position embedding,	The SQuAD dataset and the	BLEU, ROUGE-L, and METEOR evaluation metrics are used. And generates higher-order	[38]

Network for Question Generation. (2019)	self-attention units, multiple layers of gated CNN, bidirectional attention block, and answer-related scoring function. An answer-related encoder module with a Bi-LSTM network is used. Decoder: AN attention-based LSTM decoder is used.	MS MARCO dataset	factual questions by applying a syntactic-based approach.	
12.Capturing Greater Context for Question Generation. (2019)	Encoder: Input encoding consists of a bidirectional LSTM encoder, and context encoding includes soft attention affinity matrix, column-wise max pooling, and answer maskingDecoder: BiLSTM sequence to sequence decoder with pointer generator network.	The SQuAD dataset, the MS MARCO dataset, and the NewsQA dataset.	BLEU, METEOR, and ROUGE-L evaluation metrics are used for evaluation. It generates a better quality of questions by applying a multi-step recursive attention mechanism and set new state-of-the-art results.	[39]
13.Key Phrase Extraction for Generating Educational Question-Answer Pairs. (2019)	Encoder: Pre-processing includes part of speech and named entity recognition. The binary classifier includes KPE-Class (keyphrase extraction) and contains a Bidirectional LSTM encoder based on the self-attention network. Decoder: Question generation includes a bidirectional LSTM encoder based on self-attention and LSTM based decoder.	The SQuAD dataset.	Evaluated by some education experts. It generates high-quality fact-based questions and produces a first question generation tool used by the domain expert (classroom teachers) for an educational perspective.	[52]
14.Cross-Lingual Training for Automatic Question Generation. (2019)	Encoder: Encoder with 6 identical layers having two sub-layer each, the first layer is based on a multi-head self-attention mechanism and the second is based on a position-wise fully connected feed-	The IITB Hindi Monolingual Corpus, IITB Parallel Corpus, DuReader Chinese	BLEU, ROUGE-L, and METEOR metrics are used. It generatesfactual, relevant, higher-order, more syntactic, and semantic correct questions and also develops a dataset for Hindi consisting of 6555 question-answer pairs.	[9]

	forward network. Decoder: Decoder with 6 identical layers, in addition to two sub-layers in each encoder layer, includes a third sub-layer with multi-head attention.	Dataset, News Commentary Dataset, and the SQuAD dataset.		
15.Comparative study on Question-Worthy Sentence Selection for Educational Question Generation. (2019)	Encoder: Text preprocessing includes different sentence selection strategies: Random sentence (Random), Longest sentence (Longest), Concept-Rich sentence (Concept), Concept-Type-Rich sentence (ConceptType), Most Difficult sentence (Hardest), Beginning sentence (Beginning), Centroid Based Important sentence (LexRank), and Maximum Marginal Relevance Based Important sentence (MMR) and LSTM encoder. Decoder: The question construction model is based on an attention-based sequence-to-sequence learning framework using an LSTM decoder.	The SQuAD dataset.	An automatic evaluation metric includes BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, and ROUGE-L. Human Evaluation recruits three native speakers of English to the question's quality based on grammaticality, clarity, and usefulness. Here, LexRank, a sentence selection approach gives the robust performance on the TriviaQA, and MCTest datasets (non-educational) and RACE, and LearningQ datasets (educational) by calculating eigenvector centrality in a graph of sentence similarities, and educational questions feature a more diverse set of source sentences.	[37]
16. Learning to generate questions by learning what not to generate. (2019)	Encoder: The masked feature-rich encoder is based on the bidirectional Gated Recurrent Unit (BiGRU), taking word vector, answer position, lexical and binary features, clue indicator, and word frequency features. Decoder: The decoder is based on the Gated Recurrent Unit with the attention and	The SQuAD dataset, the NewsQA dataset.	The evaluation results are reported by BLEU, ROUGE-L, and METEOR metrics by comparing them with some baselines. It outperforms BLEU-4 (17.55), ROUGE-L (44.53), and METEOR (21.24) instead of previous state-of-the-art 16.17, 44.24, 19.67 respectively on the SQuAD dataset. And BLEU-4 (11.06), ROUGE-L (43.16), and METEOR (17.43) instead of previous state-of-the-art 9.94, 42.25, 16.72	[53]

	copying mechanism.		respectively on the NewsQAdataset.This study alleviates the issue of a one-to-mapping problem by focussing on what to ask and how to ask and generating answer-aware questions by proposing a GCN-based clue prediction model and outperforms the ranges of existing state-of-the-art approaches.	
17.Neural Question Generation: Transformed (2019)	This transformer-based model includes Du et al model, BERT-fused NMT model and GPT-2 model for finetuning the existing models.	The SQuAD dataset	BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR and ROUGE-L metrics are used.	[65]
18.Question Generation by Transformers (2019)	This transformer-based model includes training on inverted dataset.	The SQuAD dataset	WER (Word Error Rate) evaluation is used.	[64]
19.Self-Attention Architectures for Answer-Agnostic Neural Question Generation (2019)	This transformer-based model includes placeholder strategy, copy mechanism and contextual word embedding.	The SQuAD dataset	BLEU-1, BLEU-2, BLEU-3, BLEU-4 and ROUGE-L metrics are used.	[62]
20.A Recurrent BERT-based Model for Question Generation (2019)	This transformer-based model includes BERT-QG model, BERT-SQG model and BERT-HLSQG model.	The SQuAD dataset	BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR and ROUGE-L metrics are used.	[67]
21.Reinforced Multitask approach for multi-hop question generation. (2020)	Encoder: Answer Encoding with answer tagging features, hierarchical document encoding with answer encoding feature and a bi-directional LSTM encoder, and Multi-task learning framework for supporting facts prediction. Decoder: LSTM network with global attention and copy mechanism.	Multi-hop question answering dataset, HotPotQA	BLEU, METEOR, ROUGE-L, and SF Coverage evaluation metrics are used. It generates high-quality textual, fact-based, and higher-order reading multi-hop questions. Proposed multi-task learning improves the QG performance from 19.55 to 20.64 on BLEU-4 and the supporting fact information further improves the QG performance from 20.64 to 21.28 on BLEU-4.	[18]

22.Towards Automatic Generation of Questions from Long Answers. (2020)	The transformer-based network includes a transformer with a copy mechanism, Transformer iwslt de en model, Transformer wmten de big model, Transformer wmtenfr big model. Multi-Source Transformer consists of Multi-Source Transformer (Summary), Multi-Source Transformer (First Sentence).	The GoogleNQ dataset, the SQuAD dataset	BLEU, ROUGE-L, and METEOR evaluation metrics are used and show that the transformer-based network's performance is much better than LSTM-based models. And their best model also faces degradation in performance while using increased sentence length.	[54]
23.Stronger Transformers for Neural Multi-Hop Question Generation (2020)	The transformer-based Neural Network includes Graph-Augmented Transformer Encoder, Auxiliary Constrastive Objective for training loss and data filtering.	HotPotQA dataset	BLEU, ROUGE-L, and METEOR evaluation metrics are used. Strong transformer-based models outperform the golden approaches by a margin of 5 BLEU points.	[63]

Table 6: A brief review of the Multiple-choice Question Generation

Topic	Question generation	Distractor selection	Evaluation	Reference
1. Automatic MCQ from Text (2018)	Pre-processing includes Text Normalization, Structural analysis, Sentence Simplification, Lexical analysis, Statistical analysis, Syntactic analysis, Coreference resolution, and Word Sense Disambiguation. Sentence Selection approaches include Sentence length, Occurrence of a particular word, Part-of-speech information, Parse information, Semantic information, Machine Learning,	Distractor selection approaches include Parts-of-speech information, Frequency, WordNet, Domain Ontology, Distributional hypothesis, Pattern matching, and Semantic analysis.	Semi-automatic evaluation is used to measure the quality of cloze items, Expert language teachers or evaluators having domain knowledge are hired to measure the sentence length, simplicity or difficulty level, domain relevance, over-informative or under-informative, distractors are good, Efficiency of the system, English proficiency and	[25]

	and summarization. Key-selection approaches include: Frequency count, Parts-of-speech and Parse information, Semantic information, Pattern matching, and Machine Learning. Question Formation approaches include: By appropriate WH-word selection, Subject-Verb-Object and their relationships, Knowledge in a sentence, dependency-based pattern, syntactic transformation, Discourse connectives, Semantic information based, Semantic-based transformation.		item analysis, and quality of the question.	
2. Automatic generation of fill-in-the-blank questions with corpus-based distractors for e-assessment to enhance learning (2019)	First, the Stanford Parser and sentence length are used to select simple and worthy sentences. Then, informative sentences are selected using a coarse-grain part-of-speech tagset. Answer-key is identified by the pointwise mutual information (PMI) association technique. The first preference is given to the multiword key as compared to the unigram key. A keyword matching approach is applied to extract answer-key. In the selected sentence, the answer-key is replaced with the blank to generate the question or stem.	A pattern search approach is applied to generate the list of distractors from the corpus.	Five evaluators are employed to evaluate the generated questions.	[26]

5. Research Gap

The detailed literature survey presented above clearly highlights the importance of the problem of AQG and the vast amount of effort devoted to the problem by various researchers. Considerable progress has been made in the area especially in the last few years with the development of Machine Learning methodologies. A better understanding of the issues involved in the solution of the AQG problem is now available. However, the study of the literature indicates that the problem is far from solved and lot more needs to be done. Most of the work till date has focused on the generation of a very limited type of questions viz., MCQs and Wh-questions. Even these questions that are generated are sometimes meaningful and sometimes not so meaningful.

In this work, it is proposed to extend the frontiers of know-how in the domain of AQG. Efforts would be made towards development of approaches that yield better results that can be utilized by students and teachers. The specific objectives towards this end are outlined in the next section.

6. Proposed Objectives and Methodology

Objectives of the proposed study are to:

1. The current Neural Networks based AQGs utilize Word Embedding models to generate the plausible sections of the text on which questions can be based and then generate the questions appropriately. Efforts can be made to extend the Word Embedding model with domain specific and area specific inputs for selected domains.
2. The extended model developed in Objective 1 above would be utilized for generation of more appropriate and crisp questions using the prevalent NN architecture for AQG e.g. using a Sequence - to - Sequence NN model.
3. Transformer models have been developed very recently and have been shown to provide better results than the other approaches reported in the literature. Efforts would be made to utilize the extended Word Embedding scheme developed along with the Transformer based models towards design and development of better AQG systems.
4. Design and development of a hybrid AQG system by leveraging the transformer-based model with the seq-to-seq model to train the bigger models by utilizing the modern GPU based architectures that provide massive parallelization.

5. Develop the question evaluation technique to distinguish the logical, syntax, and semantical and educational questions from a large bunch of questions. For example, n-gram based metrics, readability indexes, sentence mover's similarity-based models and hybrid metrics developed in this work will be used to develop new evaluation criteria for determining the quality of the generated questions.

7. References

[1] M. T. H. Chi, S. A. Siler, H. Jeong, T. Yamauchi and R. G. Hausmann, "Learning from human tutoring,"*Cognitive science,* vol. 25, p. 471–533, 2001.

[2] A. Corbett and J. Mostow, "Automating comprehension questions: lessons from a reading tutor," in *Proceedings of the 1st Workshop on Question Generation*, 2008.

[3] R. E. DiPaolo, A. C. Graesser, D. J. Hacker and H. A. White, "Hints in human and computer tutoring,"*The design of instruction and evaluation: Affordances of using media and technology,* vol. 155, 2004.

[4] C.-Y. Lin, "Automatic question generation from queries," in *Workshop on the question generation shared task*, 2008.

[5] T. Marciniak, "Language generation in the context of yahoo! answers," in *online Proceedings of 1st Question Generation Workshop*, 2008.

[6] I. Kollia and G. Siolas, "Using the ibmwatson cognitive system in educational contexts," in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2016.

[7] H. Zhu, L. Dong, F. Wei, W. Wang, B. Qin and T. Liu, "Learning to ask unanswerable questions for machine reading comprehension,"*arXiv preprint arXiv:1906.06045,* 2019.

[8] X. Du and C. Cardie, "Harvesting paragraph-level question-answer pairs from wikipedia,"*arXiv preprint arXiv:1805.05942,* 2018.

[9] V. Kumar, N. Joshi, A. Mukherjee, G. Ramakrishnan and P. Jyothi, "Cross-lingual training for automatic question generation,"*arXiv preprint arXiv:1906.02525,* 2019.

[10] F. Bu, X. Zhu, Y. Hao and X. Zhu, "Function-based question classification for general QA," in *Proceedings of the 2010 conference on empirical methods in natural language processing*, 2010.

[11] I. R. Assaly and O. M. Smadi, "Using Bloom's Taxonomy to Evaluate the Cognitive Levels of Master Class Textbook's Questions.,"*English Language Teaching*, vol. 8, p. 100–110, 2015.

[12] E. Riloff and M. Thelen, "A rule-based question answering system for reading comprehension tests," in *ANLP-NAACL 2000 Workshop: Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems*, 2000.

[13] P. Mannem, R. Prasad and A. Joshi, "Question generation from paragraphs at UPenn: QGSTEC system description," in *Proceedings of QG2010: The Third Workshop on Question Generation*, 2010.

[14] A. Varga, "Le an ha 2010 wlv: A question generation system for the qgstec 2010 task b," in *Proceedings of QG2010: The Third Workshop on Question Generation*, 2010.

[15] H. Ali, Y. Chali and S. A. Hasan, "Automation of question generation from sentences," in *Proceedings of QG2010: The Third Workshop on Question Generation*, 2010.

[16] M. Liu, R. A. Calvo and V. Rus, "Automatic generation and ranking of questions for critical review,"*Journal of Educational Technology & Society*, vol. 17, p. 333–346, 2014.

[17] K. Mazidi and R. Nielsen, "Linguistic considerations in automatic question generation," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2014.

[18] D. Gupta, H. Chauhan, A. Ekbal and P. Bhattacharyya, "Reinforced Multi-task Approach for Multi-hop Question Generation,"*arXiv*, p. arXiv–2004, 2020.

[19] V. Kumar, G. Ramakrishnan and Y.-F. Li, "Putting the horse before the cart: A generator-evaluator framework for question generation from text," in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 2019.

[20] T. Klein and M. Nabi, "Learning to Answer by Learning to Ask: Getting the Best of GPT-2 and BERT Worlds,"*arXiv preprint arXiv:1911.02365*, 2019.

[21] M. Wasim, W. Mahmood, M. N. Asim and M. U. Khan, "Multi-label question classification for factoid and list type questions in biomedical question answering,"*IEEE Access*, vol. 7, p. 3882–3896, 2018.

[22] X. Yao and Y. Zhang, "Question generation with minimal recursion semantics," in *Proceedings of QG2010: The Third Workshop on Question Generation*, 2010.

[23] K. Mazidi and P. Tarau, "Automatic question generation: From nlu to nlg," in *International conference on intelligent tutoring systems*, 2016.

[24] O. Keklik, T. Tuglular and S. Tekir, "Rule-Based Automatic Question Generation Using Semantic Role Labeling,"*IEICE TRANSACTIONS on Information and Systems,* vol. 102, p. 1362–1373, 2019.

[25] D. R. Ch and S. K. Saha, "Automatic multiple choice question generation from text: A survey,"*IEEE Transactions on Learning Technologies,* 2018.

[26] B. Das, M. Majumder, S. Phadikar and A. A. Sekh, "Automatic generation of fill-in-the-blank question with corpus-based distractors for e-assessment to enhance learning,"*Computer Applications in Engineering Education,* vol. 27, p. 1485–1495, 2019.

[27] M. Last and G. Danon, "Automatic question generation,"*Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* p. e1382, 2020.

[28] M. Divate and A. Salgaonkar, "Automatic question generation approaches and evaluation techniques,"*Current Science,* p. 1683–1691, 2017.

[29] N.-T. Le, T. Kojiri and N. Pinkwart, "Automatic question generation for educational applications–the state of art," in *Advanced computational methods for knowledge engineering*, Springer, 2014, p. 325–338.

[30] K. Mazidi and R. D. Nielsen, "Leveraging multiple views of text for automatic question generation," in *International Conference on Artificial Intelligence in Education*, 2015.

[31] S. Pal, T. Mondal, P. Pakray, D. Das and S. Bandyopadhyay, "Qgstec system description–juqgg: A rule basedapproach,"*Boyer&Piwek (2010),* p. 76–79, 2010.

[32] K. Mazidi, "Automatic question generation from passages," in *International Conference on Computational Linguistics and Intelligent Text Processing*, 2017.

[33] A. Chen, G. Stanovsky, S. Singh and M. Gardner, "Evaluating question answering evaluation," in *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, 2019.

[34] P. Resnik and J. Lin, "11 evaluation of NLP systems,"*The handbook of computational linguistics and natural language processing,* vol. 57, 2010.

[35] M. Divate and A. Salgaonkar, "Good-Quality Question Generation for Academic Support," in *Advanced Computational and Communication Paradigms*, Springer, 2018, p. 777–790.

[36] K. Papineni, S. Roukos, T. Ward and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002.

[37] G. Chen, J. Yang and D. Gasevic, "A comparative study on question-worthy sentence selection strategies for educational question generation," in *International Conference on Artificial Intelligence in Education*, 2019.

[38] Y. Wang, J. Zheng, Q. Liu, Z. Zhao, J. Xiao and Y. Zhuang, "Weak supervision enhanced generative network for question generation,"*arXiv preprint arXiv:1907.00607*, 2019.

[39] L. A. Tuan, D. J. Shah and R. Barzilay, "Capturing Greater Context for Question Generation,"*arXiv preprint arXiv:1910.10274*, 2019.

[40] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005.

[41] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004.

[42] E. Clark, A. Celikyilmaz and N. A. Smith, "Sentence mover's similarity: Automatic evaluation for multi-sentence texts," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[43] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger and Y. Artzi, "Bertscore: Evaluating text generation with bert,"*arXiv preprint arXiv:1904.09675*, 2019.

[44] N. Myller, "Automatic generation of prediction questions during program visualization,"*Electronic Notes in Theoretical Computer Science*, vol. 178, p. 43–49, 2007.

[45] M. Blšták and V. Rozinajová, "Automatic question generation based on analysis of sentence structure," in *International Conference on Text, Speech, and Dialogue*, 2016.

[46] X. Du, J. Shao and C. Cardie, "Learning to ask: Neural question generation for reading comprehension,"*arXiv preprint arXiv:1705.00106*, 2017.

[47] X. Du and C. Cardie, "Identifying where to focus in reading comprehension for neural question generation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.

[48] Y. Zhao, X. Ni, Y. Ding and Q. Ke, "Paragraph-level neural question generation with maxout pointer and gated self-attention networks," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[49] P. Nema, A. K. Mohankumar, M. M. Khapra, B. V. Srinivasan and B. Ravindran, "Let's Ask Again: Refine Network for Automatic Question Generation,"*arXiv preprint arXiv:1909.05355,* 2019.

[50] B. Pan, H. Li, Z. Yao, D. Cai and H. Sun, "Reinforced dynamic reasoning for conversational question generation,"*arXiv preprint arXiv:1907.12667,* 2019.

[51] X. Lu, "Learning to generate questions with adaptive copying neural networks," in *Proceedings of the 2019 International Conference on Management of Data*, 2019.

[52] A. Willis, G. Davis, S. Ruan, L. Manoharan, J. Landay and E. Brunskill, "Key Phrase Extraction for Generating Educational Question-Answer Pairs," in *Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale*, 2019.

[53] B. Liu, M. Zhao, D. Niu, K. Lai, Y. He, H. Wei and Y. Xu, "Learning to generate questions by learningwhat not to generate," in *The World Wide Web Conference*, 2019.

[54] S. K. Mishra, P. Goel, A. Sharma, A. Jagannatha, D. Jacobs and H. Daume, "Towards Automatic Generation of Questions from Long Answers,"*arXiv preprint arXiv:2004.05109,* 2020.

[55] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient estimation of word representations in vector space,"*arXiv preprint arXiv:1301.3781,* 2013.

[56] M. Heilman, "Automatic factual question generation from text," 2011.

[57] M. Flor and B. Riordan, "A semantic role-based approach to open-domain automatic question generation," in *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, 2018.

[58] A. R. Fabbri, P. Ng, Z. Wang, R. Nallapati and B. Xiang, "Template-Based Question Generation from Retrieved Sentences for Improved Unsupervised Question Answering,"*arXiv preprint arXiv:2004.11892*, 2020.

[59] G. Danon and M. Last, "A syntactic approach to domain-specific automatic question generation,"*arXiv preprint arXiv:1712.09827,* 2017.

[60] Y. Susanti, T. Tokunaga and H. Nishikawa, "Integrating automatic question generation with computerised adaptive test,"*Research and Practice in Technology Enhanced Learning,* vol. 15, p. 1–22, 2020.

[61] M. Cubric and M. Tosic, "Design and evaluation of an ontology-based tool for generating multiple-choice questions,"*Interactive Technology and Smart Education,* 2020.

[62] T. Scialom, B. Piwowarski and J. Staiano, "Self-attention architectures for answer-agnostic neural question generation," in *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*, 2019.

[63] D. S. Sachan, L. Wu, M. Sachan and W. Hamilton, "Stronger Transformers for Neural Multi-Hop Question Generation,"*arXiv preprint arXiv:2010.11374,* 2020.

[64] K. Kriangchaivech and A. Wangperawong, "Question generation by transformers,"*arXiv preprint arXiv:1909.05017,* 2019.

[65] J. Hu, "Neural Question Generation: Transformed".

[66] Y. Chen, L. Wu and M. J. Zaki, "Reinforcement learning based graph-to-sequence model for natural question generation,"*arXiv preprint arXiv:1908.04942,* 2019.

[67] Y.-H. Chan and Y.-C. Fan, "A recurrent BERT-based model for question generation," in *Proceedings of the 2nd Workshop on Machine Reading for Question Answering,* 2019.